Chapter 14. Psychology

Richard S. Sutton and Andrew G. Barto

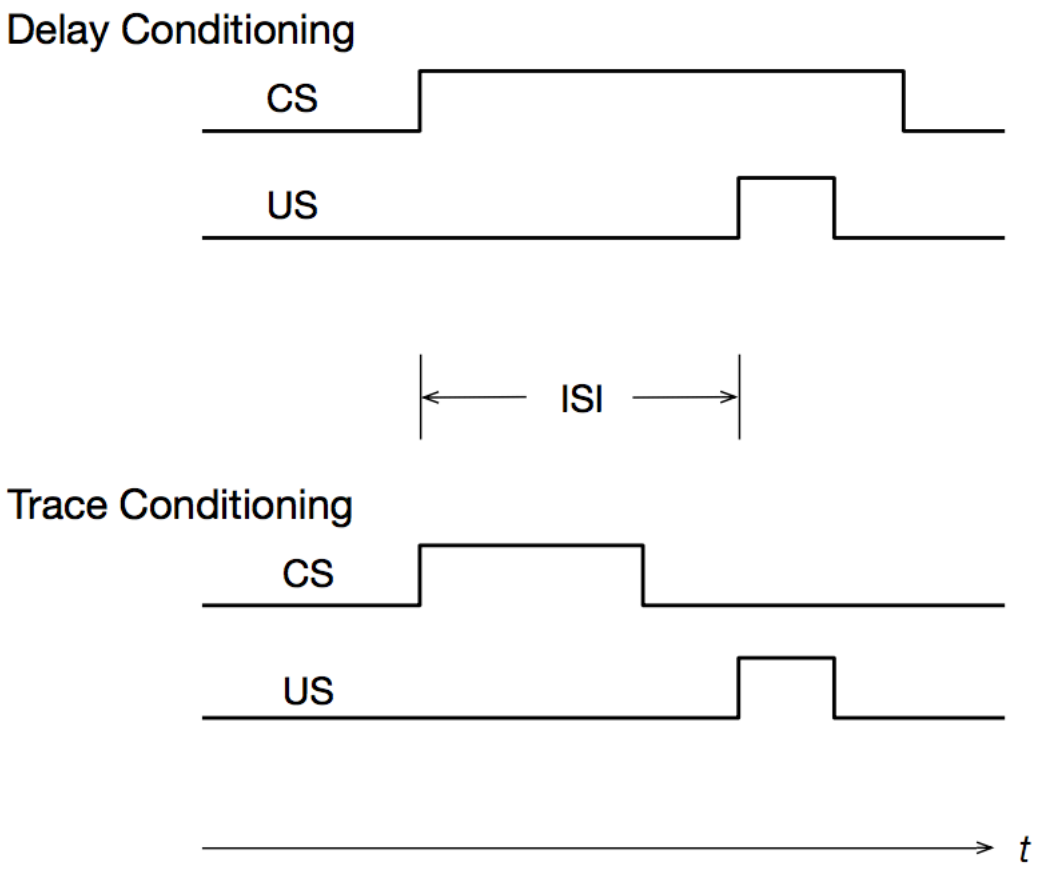
Reporter: LiuYang

14.1 Terminology

- **Reinforcement** in psychology originally referred to the strengthening of a pattern of behavior as a result of an animal receiving a stimulus. Not just to strengthening, but also to weakening a behavior pattern.
- **Reward**: Taste of nourishing food, sexual contact, successful escape. Secondary reward is the rewarding quality acquired by stimuli or events that predict primary reward or other secondary rewards. $\mathbf{R_t}$ is more like a reward signal than a reward.

14.1 Terminology

- **URs**: unconditioned responses
- **USs**: unconditioned stimuli
- **CRs**: conditioned responses
- **CSs**: conditioned stimulus



14.3 Classical Conditioning

- *It is pretty evident that under natural conditions the normal animal must respond not only to stimuli which themselves bring immediate benefit or harm, but also to other physical or chemical agencies—waves of sound, light, and the like—which in themselves only signal the approach of these stimuli; though it is not the sight and sound of the beast of prey which is in itself harmful to the smaller animal, but its teeth and claws. (Pavlov, 1927, p. 14)*

14.3.1 The Rescorla-Wagner Model

- The way to update asymptotic level of associative strength

$$\begin{aligned}\Delta V_{\mathrm{A}} &= \alpha_{\mathrm{A}}\beta_{\mathrm{Y}}(\xi_{\mathrm{Y}} - V_{\mathrm{AX}}) \\ \Delta V_{\mathrm{X}} &= \alpha_{\mathrm{X}}\beta_{\mathrm{Y}}(\xi_{\mathrm{Y}} - V_{\mathrm{AX}}),\end{aligned}$$

- ξ_Y is the asymptotic level of associative strength that the US Y can support.
- Assumption: **VAX** is equal to **VA + VX**.
- Experiments like these demonstrate. One CS component predicts a US, then learning that a newly-added second CS component also predicts the US is much reduced. This is called blocking.

14.3.1 The Rescorla-Wagner Model

- $\phi(s) = \left(\phi_1(s), \phi_2(s), \ldots, \phi_n(s)\right)^T$ where $\phi_1(s) = 1$ if CS_i is present on the trial and 0 otherwise.
- **aggregate associative strength** for trial-type s is

$$\boldsymbol{v(s, \theta) = \theta^T \phi(s)}$$

$$\boldsymbol{\theta_{t+1} = \theta_t + \alpha \delta_t \phi(S_t)}$$

- $\boldsymbol{\alpha}$ is the step-size parameter, and $\boldsymbol{\delta_t}$ is the prediction error

$$\boldsymbol{\delta_t = \xi_t - v(S_t, \theta_t)}$$

14.3.2 The TD Model

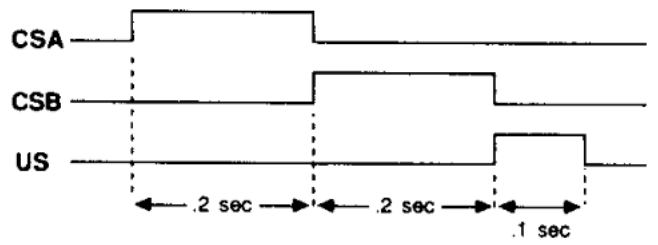
- A real-time model. Letting t label a time step, a small interval of real time, instead of a complete trial
- **Semi-gradient TD(λ)**

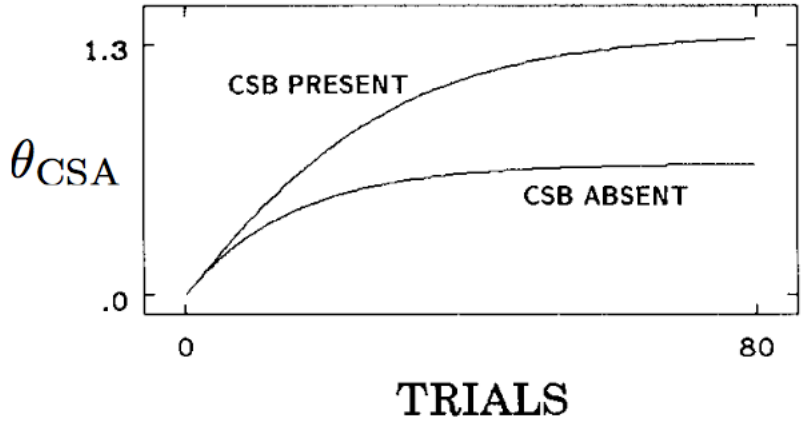
$$\boldsymbol{\theta_{t+1} = \theta_t + \alpha \delta_t e_t}$$
$$\boldsymbol{\delta_t = \xi_t + \gamma v(S_{t+1}, \theta_t) - v(S_t, \theta_t)}$$
$$\boldsymbol{e_{t+1} = \gamma \lambda e_t + \phi(S_t)}$$

- Note that if $\boldsymbol{\gamma = 0}$, the TD model essentially reduces to the Rescorla-Wagner model

14.3.3 TD Model Simulations

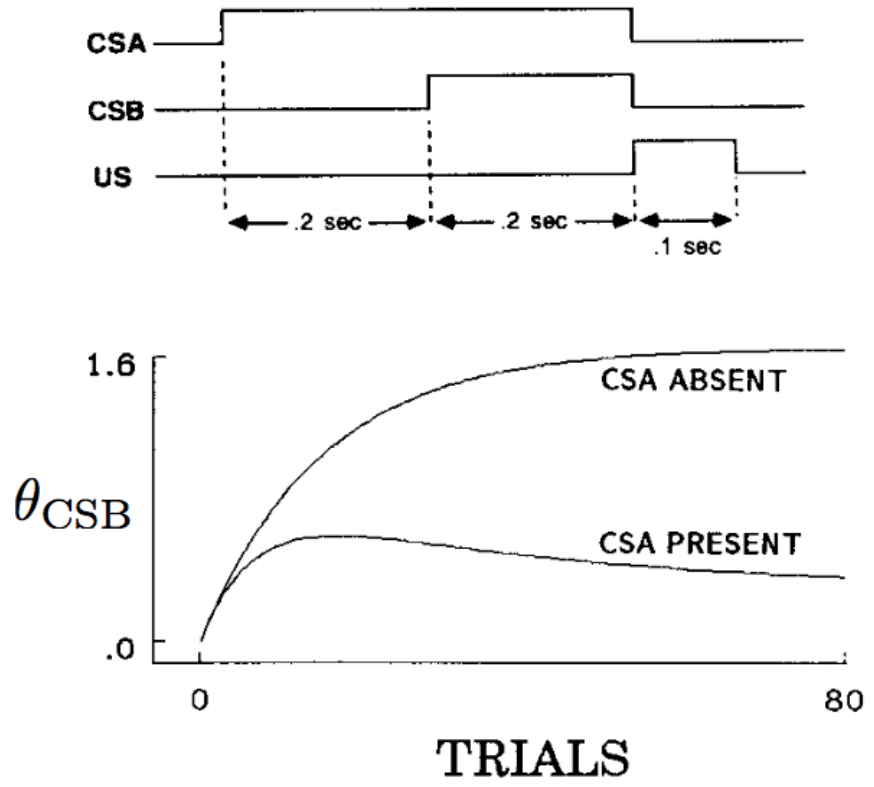
- if the empty trace interval between the CS and the US is filled with a second CS to form a serial compound stimulus, then conditioning to the first CS is facilitated.
- the model shows facilitation of both the rate of conditioning and the asymptotic level of conditioning of the first CS due to the presence of the second CS.
- Kehoe, 1982





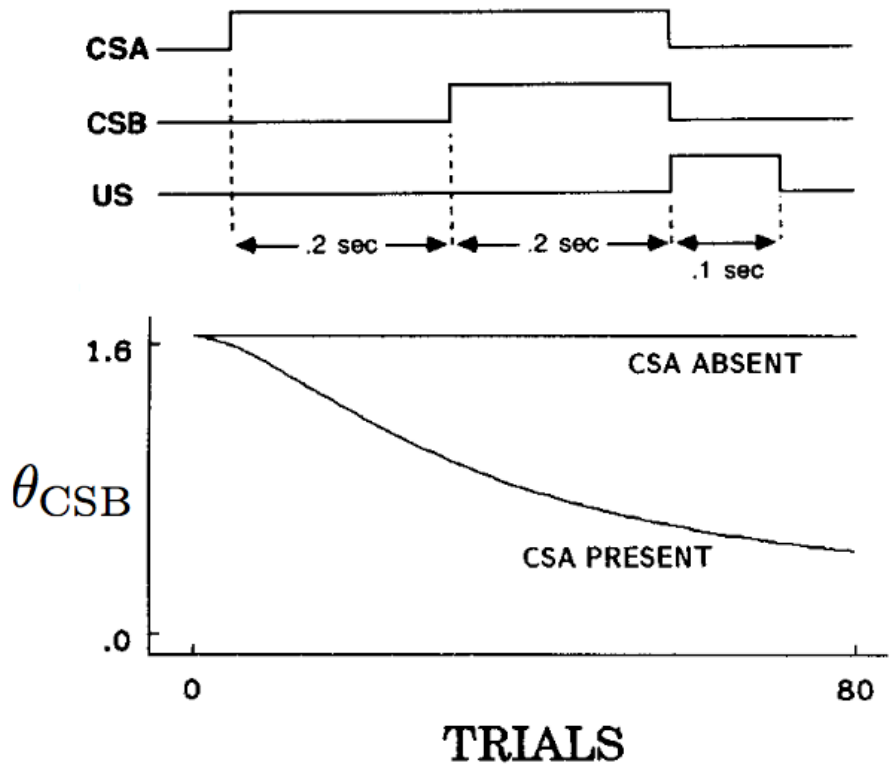
14.3.3 TD Model Simulations

- Although CSB is in a better temporal relationship with the US, the presence of CSA reduced conditioning to CSB substantially as compared to controls in which CSA was absent.
- Egger and Miller (1962)



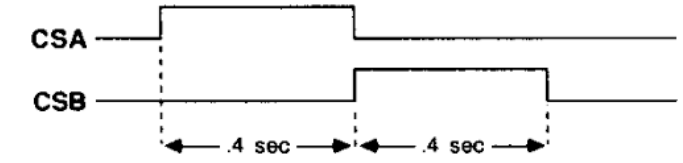
14.3.3 TD Model Simulations

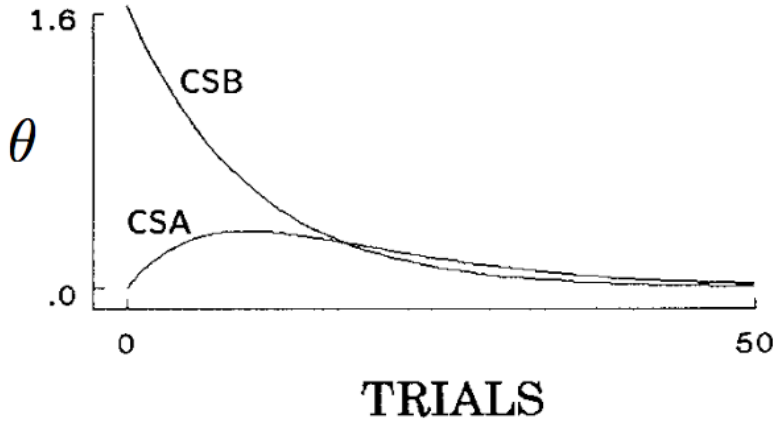
- Blocking is reversed if the blocked stimulus is moved earlier in time so that its onset occurs before the onset of the blocking stimulus.
- Recall that in blocking, if an animal has already learned that one CS predicts a US, then learning that a newly-added second CS also predicts the US is much reduced
- i.e., is blocked.
- Kehoe, Scheurs, and Graham (1987)



14.3.3 TD Model Simulations

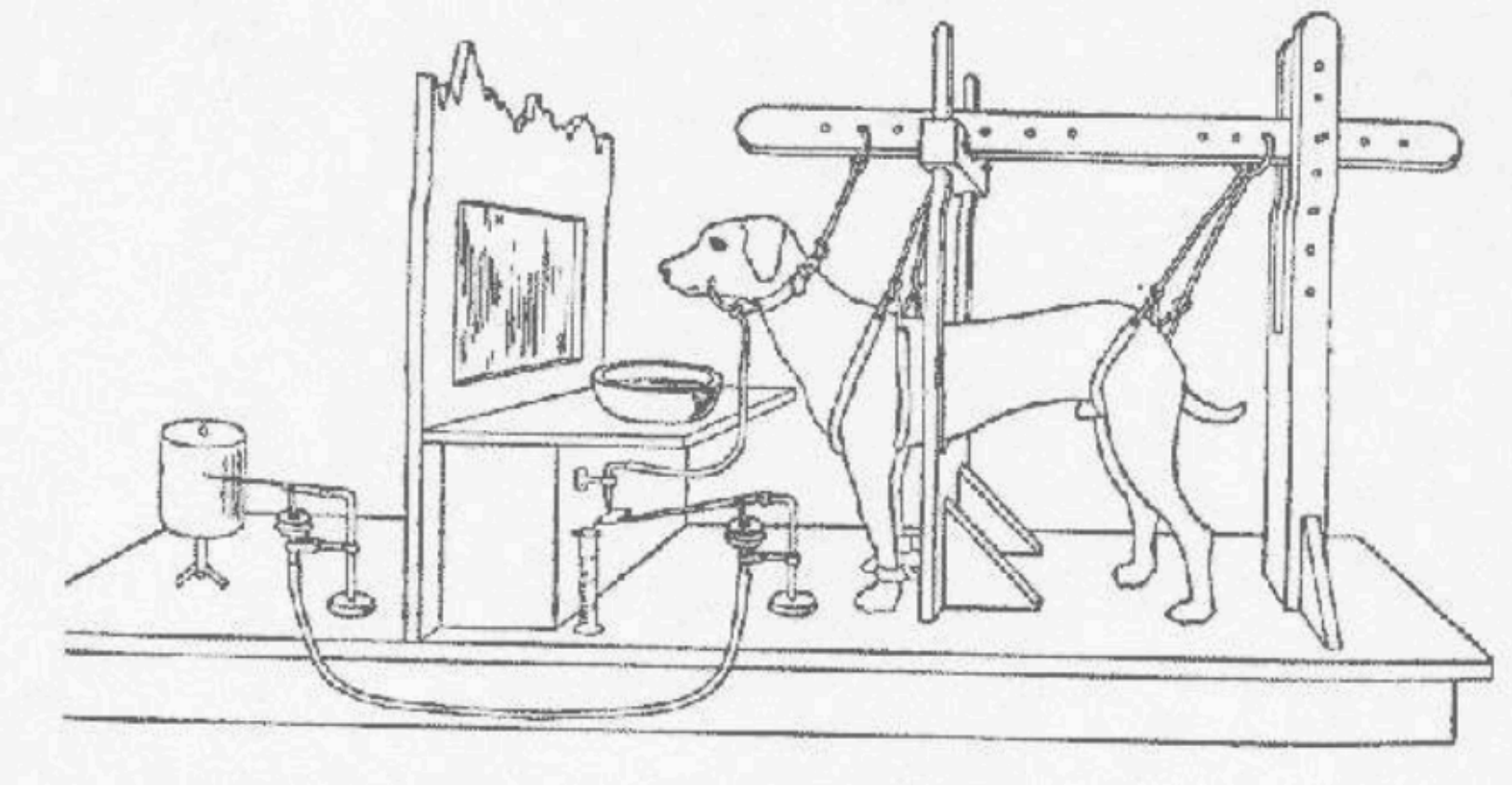
- The TD model predicts that an earlier predictive stimulus takes precedence over a later predictive stimulus.
- Second-order conditioning is the phenomenon in which a previously-conditioned CS can act as if it were a US in conditioning another initially neutral stimulus.
- Pavlov (black square experiment)





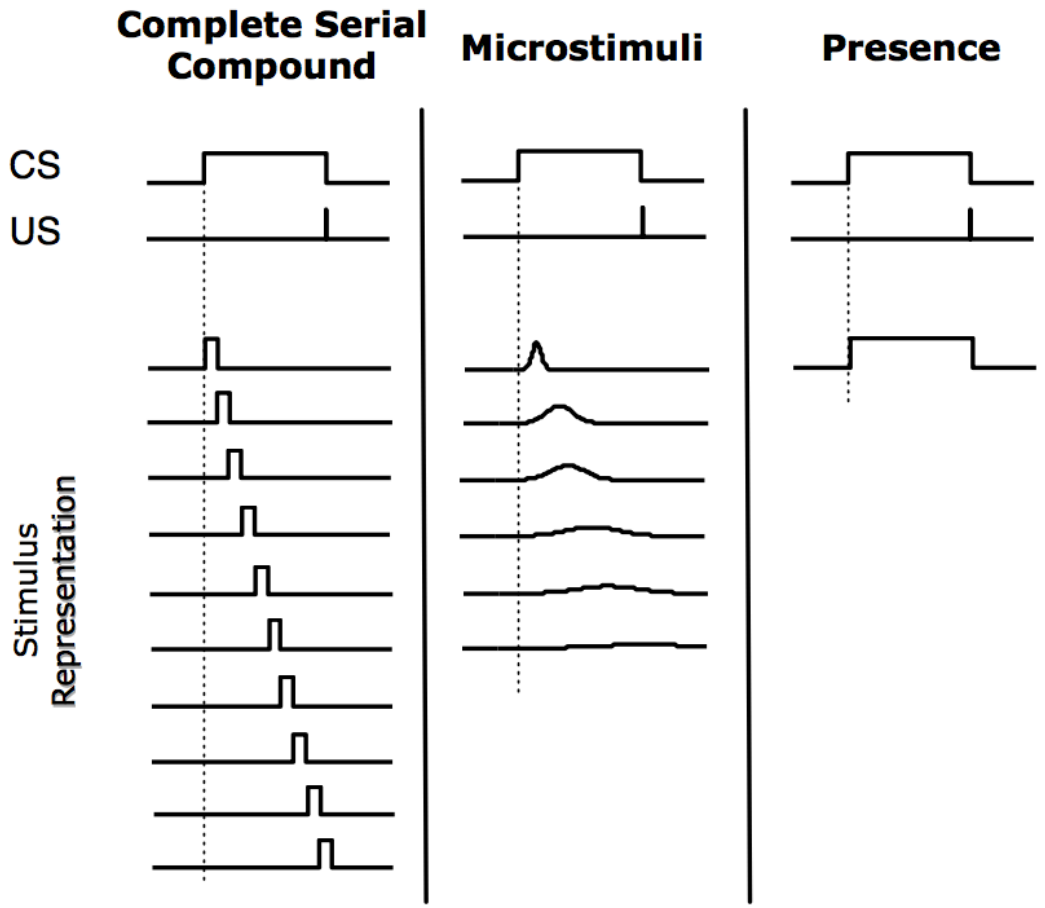
14.3.3 TD Model Simulations

- Pavlov (black square experiment)

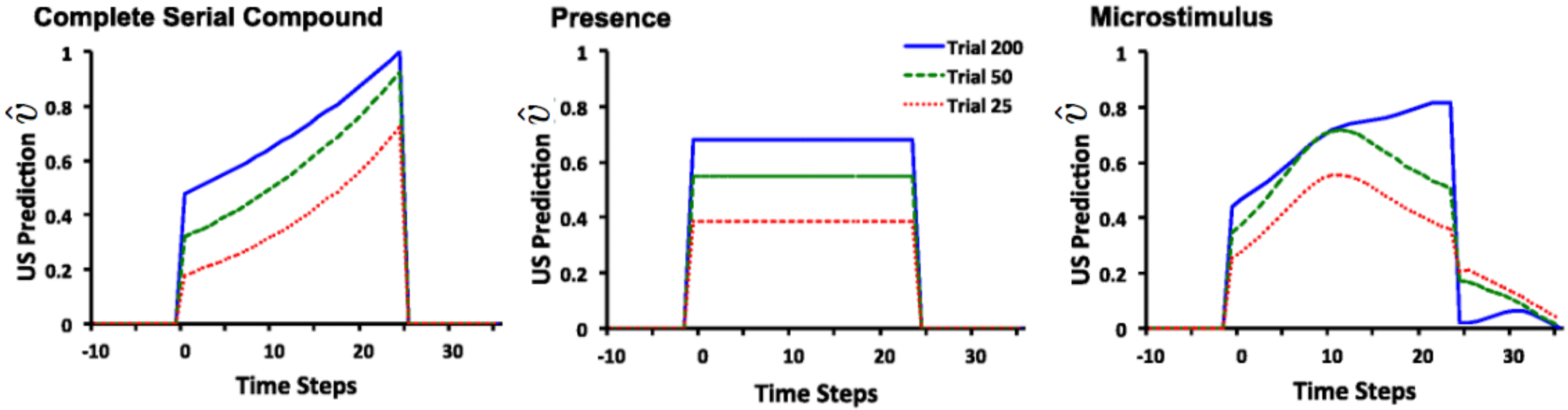


14.3.3 TD Model Simulations

- three of the stimulus representations that have been used in exploring the behavior of the TD model:
- complete serial compound (CSC)
- microstimulus (MS)
- presence representations (Ludvig, Sutton, and Kehoe, 2012).



14.3.3 TD Model Simulations



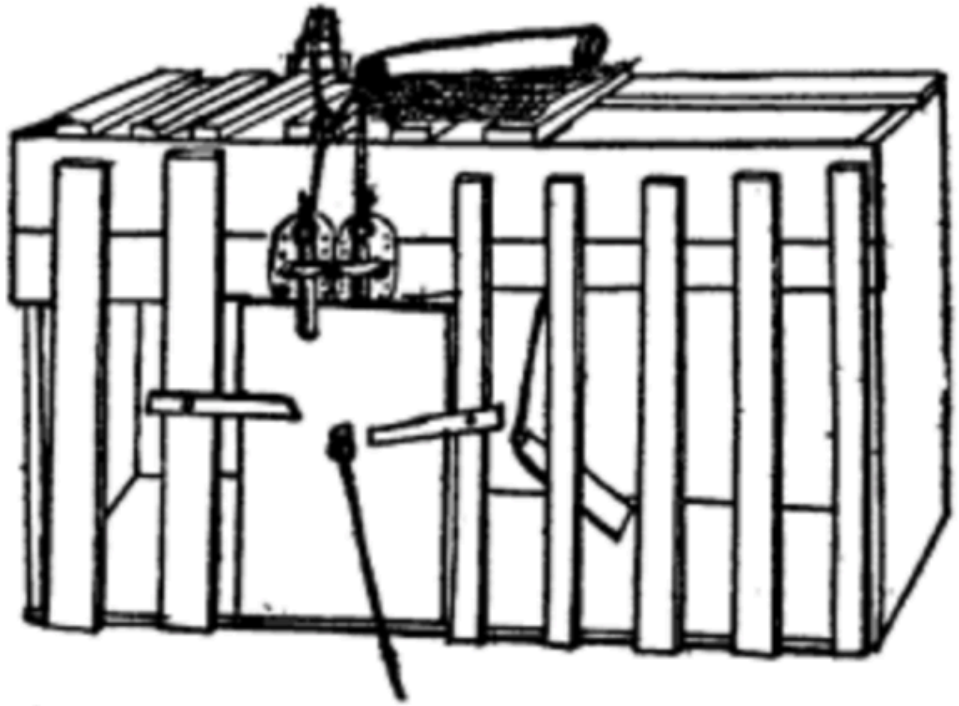
- In one trail
- US is 25 time steps after the time when CS onset. $\alpha = 0.05, \lambda = 0.95, \gamma = 0.97$
- the stimulus is present because there is only one weight (middle)

14.3.3 TD Model Simulations

- Strong influence that the **stimulus representation** has on predictions derived from the TD model.
- The TD model, when combined with particular stimulus representations, response generation mechanisms is able to account for a surprisingly-wide range of phenomena observed in animal classical conditioning experiments, but it is far from being a perfect model.
- **long-term,** instead of immediate, prediction is a key feature.

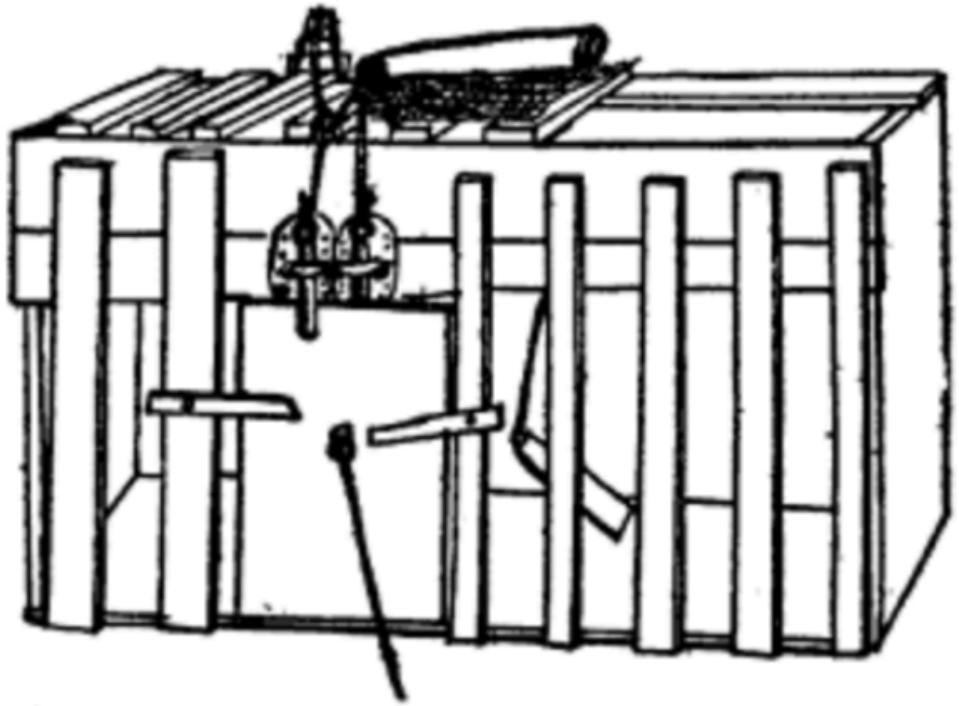
14.4 Instrumental Conditioning

- **puzzle boxes**
- search in the form of trying and selecting among many actions in each situation, and memory in the form of associations linking situations with the actions found—so far—to work best in those situations.



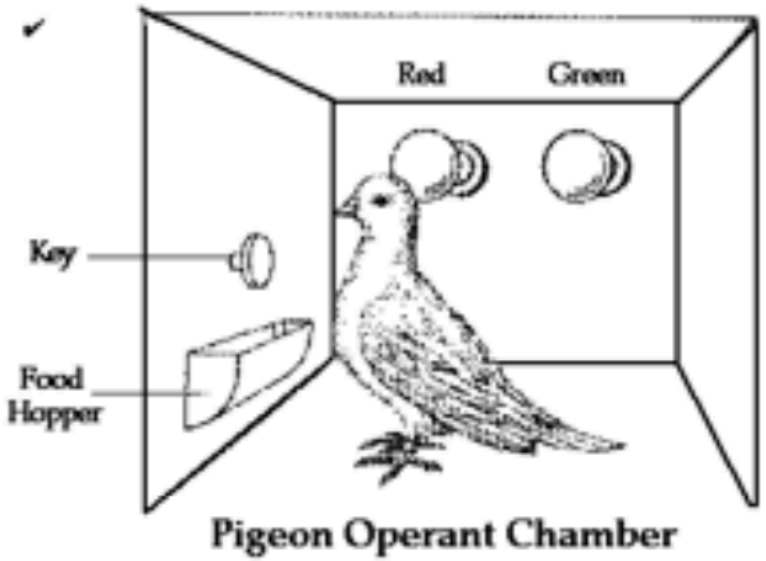
14.4 Instrumental Conditioning

- A reinforcement learning algorithm’ s need to search means that it has to explore in some way.
- ? “absolutely random, blind groping” (Woodworth, 1938, p. 777)
- demonstrate insight



14.4 Instrumental Conditioning

- **Skinner box** – the most basic version of which contains a lever or key that an animal can press to obtain a reward
- starting with an **easier problem** and incrementally increasing its difficulty as the agent learns can be an effective, and sometimes indispensable, strategy.



Pigeon Operant Chamber

14.5 Delayed Reinforcement

- Pavlov (1927) pointed out that every stimulus must **leave a trace** in the nervous system that persists for some time after the stimulus ends and proposed that stimulus traces **make learning possible** when there is a temporal gap between the CS onset and the US onset.
- the maximum strength of an instrumentally-conditioned response decreases with increasing delay of reinforcement (Hull, 1932, 1943)

14.6 Cognitive Maps

- challenged the then-prevailing stimulus-response (S–R) view
- “during the non-reward period, the rats [in the experimental group] were developing a latent learning of the maze which they were able to utilize as soon as reward was introduced” (Blodgett, 1929).

14.7 Habitual and Goal-Directed Behavior

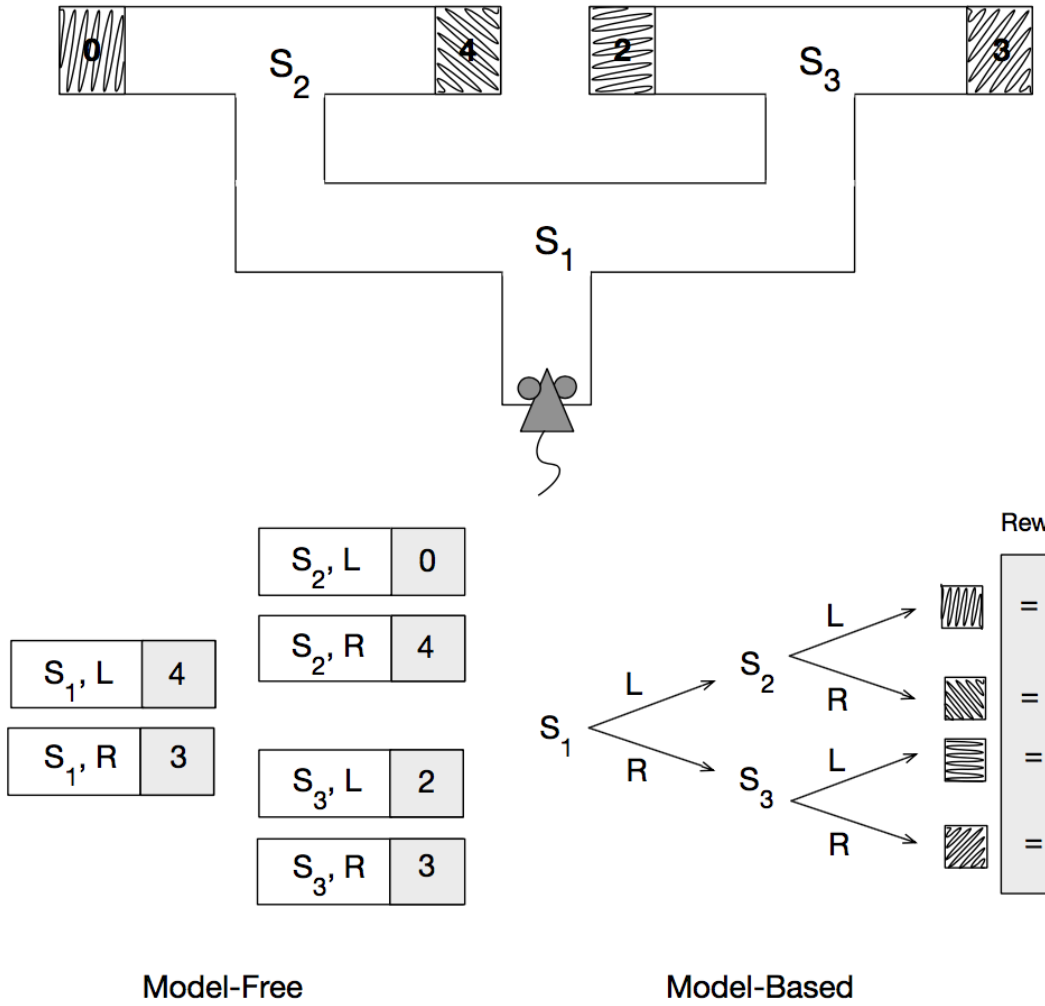
- **Habits（model free）** are behavior patterns triggered by appropriate stimuli and then performed more-or-less automatically.
- While habitual behavior responds **quickly to input** from an accustomed environment, it is **unable to quickly adjust** to changes in the environment.

14.7 Habitual and Goal-Directed Behavior

- **goal-directed behavior（model based）** is said to be controlled by its consequences (Dickinson, 1980, 1985).
- **Goal-directed control** has the advantage that it can rapidly change an animal’ s behavior when the environment changes its way of reacting to the animal’ s actions.
- The development of goal-directed behavioral control was likely a **major advance** in the evolution of animal intelligence.

14.7 Habitual and Goal-Directed Behavior

- a rat with a previously learned transition and reward model is placed directly in the goal box to the right of S2 to find that the reward available there now has value 1 instead of 4.



14.7 Habitual and Goal-Directed Behavior

- Adams and Dickinson (1981)

  Rats trained to press lever. Nausea (LiCl) after free feed (one group). More sensitive when extinction training.

- Adams (1982) conducted an experiment to see if extended training would convert goal-directed behavior into habitual behavior.

  100 vs 500 trails

14.7 Habitual and Goal-Directed Behavior

- with continued experience, the **model-free process becomes more trust-worthy** because planning is prone to making mistakes due to model inaccuracies and short-cuts necessary to make planning feasible
- one would expect a shift **from goal-directed behavior to habitual behavior** as more experience accumulates..

14.8 Summary

- Correspondences between **reinforcement learning** and the **experimental study** of animal in psychology.
- Exploration does not have to be limited to "blind groping"; trials can be generated by sophisticated methods **using innate and previously learned knowledge** as long as there is some exploration.
- The problem of delayed reinforcement: **eligibility traces** and **value functions** learned via TD algorithms.
- Environment **models** in reinforcement learning are like **cognitive maps**

Thanks. Q&A